STATISTICAL MACHINE TRANSLATION IMPROVEMENT BASED ON PHRASE SELECTION

Cyrine Nasri, Latiri Chiraz, Kamel Smaili

STATISTICAL MACHINE TRANSLATION IMPROVEMENT BASED ON PHRASE SELECTION

Cyrine Nasri
SMarT, LORIA
Campus Scientifique BP 139,
54500 Vandoeuvre Lès Nancy
Cedex, France
cyrine.nasri@loria.fr

Chiraz Latiri
LIPAH laboratory
Faculty of Sciences of Tunis
University of Tunis El Manar
Tunisia
chiraz.latiri@gnet.tn

Kamel Smaïli
SMarT, LORIA
Campus Scientifique BP 139,
54500 Vandoeuvre Lès Nancy
Cedex, France
smaili@loria.fr

Abstract

This paper describes the importance of introducing a phrase-based language model in the process of machine translation. In fact, nowadays SMT are based on phrases for translation but their language models are based on classical ngrams. In this paper we introduce a phrase-based language model (PBLM) in the decoding process to try to match the phrases of a translation table with those predicted by a language model. Furthermore, we propose a new way to retrieve phrases and their corresponding translation by using the principle of conditional mutual information.

The SMT developed will be compared to the baseline one in terms of BLEU, TER and METEOR. The experimental results show that the introduction of PBLM in the translation decoding improve the results.

1 INTRODUCTION

Language modeling is a crucial task in many areas of natural language processing (NLP) like Automatic Speech Recognition (ASR), Statistical Machine Translation (SMT), Optical Character Recognition (OCR), etc. Every improvement in the language model performance can impact the previously cited applications

Many researches on language modeling have been proposed in the literature over the past decades (Yoshua et al., 2003), (Schwenk, 2007) and (Wu et al., 2012). Nowadays, the new language models are based on deep learning techniques (Arsoy et al., 2012). Some studies were proposed to improve the language model quality by adding external informations (syntactic, morphological, etc). Significant improvements were noted (Charniak et al., 2003) (Kirchhoff and Yang, 2005) (Sarikaya and Deng, 2007) (L. Schwartz and et al., 2011), (Xiao et al., 2011).

In the following, we will be interested by variable-length models. In fact, words are commonly used as the basic lexical unit in standard language model, however in automatic speech recognition, some works were based on variable-length models where the basic unit is variable in terms of length. These variable-length ngrams correspond to phrases as defined in the speech recognition and machine translation communities. The models shown that they reduce the perplexity of the language model and sometimes they improve the performance of the ASR (Giachin, 1995) (Dietrich, 1998) (G. Riccardi and Riley, 1997) (K.F. Ries and Waibel, 1996) (Zitouni et al., 2003).

In SMT, (Baisa, 2011), first proposed the chunk-based language model (including phrase-based) in machine translation but did not give a solution. Recently, (Xu and Chen., 2015) designed a direct algorithm for phrase-based language model in statistical machine translation. In their method, phrase can be any word sequence. The phrase vocabulary is huge and the data sparsity problem is very serious. It leads to difficulty in probability estimation for phrase-based language model.

Language model is considered as the one of the most important component in SMT. Its role is to assign a probability to each translation hypothesis. In this paper, we propose to extend the standard language model to a variable-length one by considering phrases as atomic units in a language model.

This approach has the following major advantages: the first is that the phrase-based language model can easily capture a relationship between words over a long distance, within a sentence. The second advantage, is the compatibility of the translation hypotheses with that of the language model, ensuring more consistency in the decoding process. It means that we hope that the translation

hypotheses would correspond to the units of the language models.

We integrated this new language model in two statistical translation systems: baseline phrase-based SMT system (Koehn et al., 2003), and inter-lingual triggers based machine translation (Nasri et al., 2014).

This paper is structured as follows: first we give an overview of inter-lingual triggers. Second we present our method for training phrases for SMT. Then we describe our approach to derive a new phrase-based language model to be included as such a new statistical machine translation system. Finally, we present results of the proposed translation system using the new phrase-based language model. We end with a conclusion which points out the strength of our method and gives some tracks about the future work.

2 INTER-LINGUAL TRIGGERS

Inter-lingual triggers are inspired from triggers concept used in statistical language modeling (Tillmann and Ney, 1997). A trigger is a set composed of words and its best correlated triggered words in terms of mutual information (MI). In (Lavecchia et al., 2007), authors proposed to determine correlations between words belonging to two different languages. Each inter-lingual trigger is composed of a triggering source linguistic unit and its best correlated triggered target linguistic units. Based on this idea, they found among the set of triggered target units, potential translations of the triggering source words. Inter-lingual triggers are determined on a parallel corpus according to mutual information measure namely:

$$MI(a,b) = P(a,b) log \frac{P(a,b)}{P(a)P(b)} \quad (1)$$

Where *a* and *b* are respectively a source and a target words. *P(a, b)* is the joint probabilities and *P(a)* and *P(b)* are marginal probabilities. For each source unit a, the authors kept its k best target triggered units. This approach has been extended to take into account triggers of phrases (Lavecchia et al., 2008). The drawback of this method is that phrases are built in an iterative process starting from single words and joining others to them until the expected size of phrases is reached. In other words, at the end of the first iteration, sequences of two words are built, the following iteration produces phrase of three words and so on until the stop-criteria is reached. Then, once all the source phrases are built, their corresponding phrases in the target language are retrieved by using *n*-to-*m* inter-lingual trigger approach which means that a phrase of n words triggers a phrase of m words. In order to avoid the propagation of errors due to the cascade of steps in the previous method, we propose a new approach which is based on a conditional mutual information which allows retrieving target phrases given source ones.

3 A NEW METHOD FOR LEARNING PHRASE TRANSLATIONS

In this section, we present our new approach to learn a translation model based on conditional mutual information (CMI). Before presenting our approach, we introduce some necessary formalizations related to CMI.

3.1 A REVIEW OF CONDITIONAL MUTUAL INFORMATION (CMI)

In order to capture the relationship between several words at least 3, we decided to use conditional mutual information which is defined as follows for discrete random variables:

$$CMI(X,Y|Z) = \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} P(x,y,z) \; log \frac{P(x,y,z)P(z)}{P(x,z)P(y,z)} \quad (2)$$

Where P is the joint or the marginal probability depending on the number of the parameters.

We suppose that random variables *X* and *Z* and *Y* and *Z* are both independent, the preceding formula could be written as follows:

$$CMI(X,Y|Z) = \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} P(x,y,z) \; log \frac{P(x,y,z)}{P(x)P(y)P(z)} \quad (3)$$

When we would like to calculate the CMI for only 3 values which correspond to 3 words in our case, the preceding formula is rewritten as follows:

$$CMI(x,y,z) = P(x,y,z) log \frac{P(x,y,z)}{P(x)P(y)P(z)} \quad (4)$$

3.2 A NEW ALGORITHM FOR LEARNING TRANSLATION PAIRS

We describe our learning phrase translations algorithm. This algorithm does not require an initial word-to-word alignment, nor an initial segmentation of the monolingual text (Costa-Jussà et al., 2010). It uses the conditional mutual information between the source and target words to identify directly phrase pairs.

Once all phrase pairs are extracted, we segment source and target training corpus in terms of the best phrases. Then, we associate to each source phrase its best target translation.

Conditional mutual information calculates the correlation relationship between n variables. This principle is interesting since it allows to associate n words in the target language to a source phrase. Such as in Lavecchia 2008, our objective is to use the principle of inter-lingual triggers except that we use multivariate mutual information. As illustrative example, guess that we are interested by phrases of length 2 which are translated by one word. For instance, *good morning* is translated by *bonjour* in French. We can then calculate directly the correlation degree between these two linguistic units as follows:

According to formula 4, for this example x = *good*, y = *morning* and z =*bonjour*.

This formula can capture this strength relationship between the words of the source phrase and the word of the target language. In fact, the equation takes into account the relationship between each component of the source phrase and the word of the target language. We believe that this will lead to more realistic phrases with more relevant translations.

Given a sentence pair (f, e), where f is a sentence in a source language and e a sentence in a target language. First, we calculate a Source-to-Target two-to-one (Trig 2-1) trigger model since CMI permits to find triggers like $x, y \longrightarrow z$ where x, y are contiguous words on a source language and z is a word in a target language. Only the k best triggers for each source phrase are kept to be incorporated into the dictionary. Then, the source phrases of the resulting triggers are sorted in a decreasing order of the CMI value. These phrases are useful to segment the source training corpus by merging two different words into one phrase.

Once the source training corpus is segmented into phrases, we determine for each source phrase its best translations in the target language. For this, we compute a Target-to-Source two-to-one triggers model like $\langle x, y \rangle \rightarrow z$, where x, y represent words in the target language and z is a token (single word or phrase) in the source language. This process is iterated to extend the length of phrases until we reach the maximum length of phrases.

The correponding process is given in Algorithm 1. At the end of this process, we get a list of triggers of source phrases with their best phrase translations, some of them are presented in Table1.

The phrases get are used to rewrite the training corpus, Table 2 gives an overview of the obtained corpus.

Algorithm 1 A phrase model based on CMI

1: S is a source corpus and T is a target corpus.
2: Train a trigger model $2 \longrightarrow 1$ where the left phrase come from S and the right one from T. For each source phrase, only the k best ones are kept.
3: Sort the phrases (the right member of the triggers) in a decreasing order of the CMI.
4: Segment the source corpus with the source phrases.
5: Execute 2, 3 and 4 but switch the source and the target corpora.
6: Calculate triggers $1 \longrightarrow 1$ where the left sequence come from S and the right one from T.
7: Go to step 2 which will increase the size of phrases until the expected length is achieved.

4 GETTING A NEW PHRASE-BASED LANGUAGE MODELS

The role of the language model in machine translation is to measure the fluency and the well-formness of a translation. Common applications of language models include estimating the distribution based on N-gram coverage of words, to predict word and word orders (Lafferty et al., 2001) (Stolcke, 2002). In this work, we propose to model the prediction of phrase and phrase orders. By considering all word sequences as phrases, the dependency inside a phrase, is preserved. In other words, word-based language model is a special case of phrase-based language model if only single word phrases are considered. Intuitively our approach has the following advantages:

Source phrases	Target phrases	CMI
parlement+européen	european+union	**0.65**
	the parlement+européen	0.52
	parlement	0.5
	européen	0.31
prendre+en+considération	bare+in+mind	**0.42**
	consider	0.32
	take+into+account	0.25
je+voudrais+remercier	I+want+to+thank	**0.62**
	I+thank	0.35
	thank+you	0.11

Table 1: Example of interlingual phrases

we must bare+in+mind the community as+a+whole. nous devons prendre+en+considération la communauté dans+son+ensemble.
mr+president I wish+to+congratulate mrs+poulen on her report. monsieur+le+président je tiens+à+féliciter madame+poulen sur+son+rapport.
madam+president the last+week the mep karla+peijs was attacked in brussels. madame+la+présidente la semaine+dernière le mep karla+peijs a été attaqué à bruxel.
you have requested a+debate+on+this+subject in the course of the+next+few+days during this part+session. tu as demandé un+débat+sur+ce+sujet au cours des+prochains+jours au cours de cette+partie +de+session.

Table 2: Example of sentences in the training corpus

Source	il faut prendre en considération le fait que les compagnies d'assurance ont besoin d'un certain temps.
Baseline	it must be taken into account the fact that insurance companies need some time.
Interlingual Triggers	Account must+be+taken of the+fact that insurance companies request a certain amount of time.
Interlingual Triggers + PBLM	we must bare+in+mind the+fact that insurance companies need some time.
Source	Dans ce contexte, il faut veiller, si une partie à l'accord opère au niveau régional.
Baseline	In this connection, we have to make sure that if the party to an agreement operates at regional level, .
Interlingual Triggers	In+this+context, it+must+be+ensured, if a party to the agreement operates at regional+level.
Interlingual Triggers + PBLM	In+this+context, we+have+to+make+sure, if a party to the agreement operates at regional+level.

Table 3: Few examples of translations based on the phrase-based language model

- To take into account long distance dependency: the phrase-based language model can easily capture the long distance relationship between the different components of the sentence.

- To ensure a consistency between phrases of the language model and those of the translation table: Considering the pertinent phrases as single units will reduce the entropy of the language model. More importantly, the current statistical machine translations are performed on phrases, which are considered as translation units. The objective is to ensure that the translated segment correspond to the phrase predicted by a language model.

To build the new phrase-based language model (PBLM), we use a segmented target training corpus in terms of phrases. It consists of 600.000 sentences extracted from the European parliament corpus Europarl. The segmentation has been achieved by using the phrases of translation, as described in the previous section. To train the model, we use SRILM (Stolcke, 2002) to build a 5-gram language model.

5 EXPERIMENTAL EVALUATION AND RESULTS

This section describes the performance of the proposed language model in a machine translation task. The system used in this test is based upon MOSES, briefly described in (Koehn et al., 2007). The parallel corpus used for training consists of French, English text from Europarl Parliament proceeding corpus (Europarl) version 6 described in Table 4. In the baseline phrase-based SMT system four models have been used, namely: four models namely: a translation table, a language model, a distortion model and a penality which reflects the difference in size between the proposed translation and the sentence to be translated. To estimate the optimal value of each weight, the Minimum Error Rate Training (MERT) algorithm is used on a development corpus. In this work, we assume that the maximum size of a phrase is 8 words. In (Nasri et al., 2014), the authors showed that the quality of translation does not increase with phrase size greater than 8 words. The development and test corpus must be rewritten in the same way as the training corpus with phrases. In case of conflict between two phrases, the algorithm will prefer the phrase with the highest CMI value.

Corpus		**French**	**English**
Training	**Sentences**	1M	
	Words	23362869	20498748
	Vocabulary	968081	967065
Dev	**Sentences**	1400	
	Words	38741	34839
Test	**Sentences**	500	
	Words	5.8k	5.3k

Table 4: Description of Europarl corpus

In this evaluation, we compare the performance of the following translation systems in terms of BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2005) and TER (Snover et al., 2006): the baseline translation system (Koehn et al., 2007) using a standard ngram language model, and the inter-lingual trigger based translation system (Nasri et al., 2014) using both models (ngram and phrase-based language model). Table 3 shows some examples of translations based on the phrase-based language model. Table 4, 5 and 6 present respectively the results in terms of BLEU, METEOR and TER.

System	Dev	Test
Baseline	30.42	28.56
Baseline + PBLM	30.76	28.8
Triggers	28.58	26.66
Triggers + PBLM	29.60	27.54

Table 5: Evaluation of translation systems using different LM (ngram PBLM) in terms of BLEU

System	Dev	Test
Baseline	50.22	49.32
Baseline + PBLM	50.91	49.61
Triggers	48.31	47.03
Triggers + PBLM	48.42	47.21

Table 6: Evaluation of translation systems using different LM (ngram PBLM) in terms of METEOR

The Phrase-Based Language Model (PBLM) while outperforms slightly the translation quality of the baseline phrase-based SMT system what-

System	Dev	Test
Baseline	35.32	30.59
Baseline + PBLM	35.24	30.29
Triggers	38.68	32.33
Triggers + PBLM	38.51	32.21

Table 7: Evaluation of translation systems using different LM (ngram PBLM) in terms of TER

ever the measures. In fact, in terms of BLEU the improvement is equal to 0.34% on Dev2010, and 0.24% on test2010. In terms of METEOR, an increase of 0,69% and 0.29% have been achieved on DEV20110 and Test2010. While for the TER we observed a reduction of TER of 0,08 and 0,12 on respectively DEV2010 and Test2010. In trigger-based machine translation, the PBLM improves also the translation quality measured by BLEU, METEOR and TER. In term of BLEU, the improvement is equal to 1.02% on Dev2010, and 0.88% on test2010. METEOR also increased of 0.11% on Dev2010 and 0.18% on test2010. TER decreased of 0,17% and 0,12% on respectively DEV2010 and Test2010.

6 CONCLUSION

In this paper, we have presented a new phrase-based language model for statistical machine translation. We first, gave the definition of interlingual triggers. Then, we described a new algorithm for learning translation pairs without an initial word-to-word alignments, nor an initial segmentation of the monolingual text. Finally, we designed a new phrase based langue model.
The experiments on French-to-English translation demonstrated that the proposed phrase-based language model improve the quality of translation by proposing another kind of language model. In fact, a variable-length language model has the ability to use potentially the same phrases as those of the partial translations which reinforces the quality of translation.

References

E. Arsoy, T. N. Sainath, B. Kingsbury, and B. Ramabhadran. 2012. Deep neural network language models. In *In Proceedings of NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 20–28.

V. Baisa. 2011. Chunk-based language model and machine translation. Master's thesis.

E. Charniak, K. Knight, and K. Yamada. 2003. Syntax-based language models for statistical machine translation. In *In MT Summit IX. Intl. Assoc. for Machine Translation*, pages 40–46.

M. R. Costa-Jussà, V. Daudaravicius, and R. E. Banchs. 2010. Using collocation segmentation to extract translation units in a phrase-based statistical machine translation system. *Procesamiento del Lenguaje Natural*, 45:215–220.

K. Dietrich. 1998. Language-model optimization by mapping of corpora. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 701–704, Seattle, USA. International Conference on Acoustics, Speech, and Signal Processing.

A. Ljolje G. Riccardi, A. L. Gorin and M. Riley. 1997. A spoken language system for automated call routing. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 1143–1146, Munich, Germany. International Conference on Acoustics, Speech, and Signal Processing.

E. Giachin. 1995. Phrase bigrams for continuous speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 225–228, Detroit, Michigan, USA. International Conference on Acoustics, Speech, and Signal Processing.

D. Buo K.F. Ries and A. Waibel. 1996. Class phrase models for language modeling. In *Proceedings of the International Conference on Spoken Language Processing*, pages 398–401, Philadelphia, USA. The Fourth International Conference on Spoken Language Processing.

K. Kirchhoff and M. Yang. 2005. Improved language modeling for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 125–128, Ann Arbor, Michigan. Association for Computational Linguistics.

P. Koehn, F.J Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Edmonton, Canada. Association for Computational Linguistics.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180,

Prague, Czech Republic. Association for Computational Linguistics.

C. Callison-Burch L. Schwartz and, W. Schuler, and S. Wu. 2011. Incremental syntactic language models for phrase-based translation.

J.D. Lafferty, A. McCallum, and F.C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

C. Lavecchia, K. Smaili, D. Langlois, and J.P Haton. 2007. Using inter-lingual triggers for machine translation. In *INTERSPEECH*, pages 2829–2832. ISCA.

C. Lavecchia, D. Langlois, and K. Smaili. 2008. Discovering phrases in machine translation by simulated annealing. In *INTERSPEECH*, pages 2354–2357. ISCA.

A. Lavie and A. Agarwal. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. pages 65–72.

C. Nasri, K. Smaili, and C. Latiri. 2014. Training phrase-based smt without explicit word alignment. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 233–241, Nepal.

K. Papineni, S. Roukosand W. Todd, and Z. Wei-Jing. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Sarikaya and Y. Deng. 2007. Joint morphological-lexical language modeling for machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Prague, Czech Republic. Association for Computational Linguistics.

H. Schwenk. 2007. Continuous space language models. *Computer Speech and Language*, 21(3):492 – 518.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

A. Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *SRILM - An Extensible Language Modeling Toolkit*, pages 901–904.

C. Tillmann and H. Ney. 1997. Word triggers and the em algorithm. In *IN PROCEEDINGS OF THE WORKSHOP COMPUTATIONAL NATURAL LANGUAGE LEARNING (CONLL 97*, pages 117–124.

Y. Wu, X. Lu, H. Yamamoto, S. Matsuda, C. Hori, and H. Kashioka. 2012. Factored language model based on recurrent neural network.

T. Xiao, J. Zhu, and M. Zhu. 2011. Language modeling for syntax-based machine translation using tree substitution grammars: A case study on chinese-english translation. pages 18:1–18:29.

J. Xu and G. Chen. 2015. Phrase based language model for statistical machine translation. arXiv preprint arXiv:1501.04324.

B. Yoshua, D. Réjean, V. Pascal, and J. Christian. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

I. Zitouni, K. Smaili, and J.P. Haton. 2003. Statistical language modeling based on variable-length sequences. volume 17, pages 27–41.